Sirius_Translators at OSACT6 2024 Shared Task: Fin-tuning Ara-T5 Models for Translating Arabic Dialectal Text to Modern Standard Arabic

Salwa Alahmari[1,2], Eric Atwell[1] and Hadeel Saadany[3],

[1]University of Leeds, UK, [2]University of Hafr Al Batin, Saudi Arabia, [3]University of Surrey, UK
scssala@leeds.ac.uk, E.S.Atwell@leeds.ac.uk, hadeel.saadany@surrey.ac.uk

Abstract

This paper presents the findings from our participation in the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT6) in 2024. Our specific focus was on the second task (Task 2), which involved translating text at the sentence level from five distinct Dialectal Arabic (DA) (Gulf, Egyptian, Levantine, Iraqi, and Maghrebi) into Modern Standard Arabic (MSA). Our team, Sirius_Translators, fine-tuned four AraT5 models namely; AraT5 base, AraT5v2-base-1024, AraT5-MSA-Small, and AraT5-MSA-Base for the Arabic machine translation (MT) task. These models were fine-tuned using a variety of parallel corpora containing Dialectal Arabic and Modern Standard Arabic. Based on the evaluation results of OSACT6 2024 Shared Task2, our fine-tuned AraT5v2-base-1024 model achieved an overall BLEUscore of 21.0 on the development (Dev) set and 9.57 on the test set, respectively.

1 Introduction

To emphasize the significance of addressing Arabic dialects, it's noteworthy that Ethnologue[1] ranks Arabic as the language with the 5th highest number of native speakers, totalling approximately 420 million individuals across 22 countries in the Middle East and North Africa region. Arabic is characterized by diglossia, a linguistic phenomenon where MSA is used in formal contexts, while DA is prevalent in informal settings (Al-Sobh et al., 2015; Abdul-Mageed et al., 2022). Dialects are broadly categorized by region, such as Egyptian or Gulf dialects, but they also exhibit nuanced variations even within individual countries. The linguistic variation presents substantial challenges for MT models trained on MSA. Employing these models, designed specifically for MSA, on DA can be problematic, resulting in subpar translation outcomes when applied to DA. One potential solution to overcome this challenge involves creating parallel corpora, including MSA translations of text written in DA. Recently, considerable efforts have been devoted to translating dialects into MSA. However, the prevalent approach across most studies involves treating each dialect independently. As a result, it is crucial to formulate models with the capability to collectively manage and process at least the most common Arabic dialects.

In this paper, we detail the experiments conducted to develop DA MT model. More precisely, we evaluate the results of fine-tuning different architectures (versions) of the AraT5 transformer model (Nagoudi et al., 2021), employing various datasets for the training phase. The structure of the paper is as follows: Section 2 provides background information about Arabic dialects. Section 3 outlines related works. Section 4 describes the dataset used. The research methodology, including the fine-tuning of AraT5 models and training configuration, is presented in Section 5. In Section 6, we discuss the obtained results. Finally, Section 7 offers a conclusive summary and discusses potential future work.

2 Arabic Dialects Overview

Provided here is contextual background on the variation found in Arabic dialects. MSA represents the formal variant of Arabic, taught in educational institutions and utilized for formal texts and news presentations. MSA has its roots in the Classical Arabic of the Qur'an, albeit experiencing changes in vocabulary and specific aspects of grammar over time. Nevertheless, the majority of Arabs speak their regional dialect as their natural language which is notably different from MSA form of Arabic. While the precise categorization of regional dialects may not be entirely consistent, here are a few main groups:

[1]https://www.ethnologue.com

OSACT 2024 Workshop, pages 117–123

1. **Gulf:** spoken in Gulf countries including Saudi Arbia, Kuwait, Bahrain, Oman and Qatar.
2. **Egyptian:** spoken in Egypt only.
3. **Levantine:** spoken in Levant countries including Lebanon, Jordan, Syria, and Palestine.
4. **Iraqi:** spoken in Iraq and regions of neighbouring countries, also referred to as Mesopotamian Arabic.
5. **Maghrebi:** Spoken in Morocco, Algeria, Tunisia, Libya, Western Sahara, and Mauritania, Maghrebi is influenced by French and Berber(Turki et al., 2016)

Elaborating on how dialectal variations may manifest in their written form, is detailed from a Natural Language Processing(NLP) perspective by Zaidan and Callison-Burch (2014). For example, concerning morphology, they observe that the absence of grammatical cases in dialects is primarily evident in the accusative when a suffix is introduced. This is attributed to the fact that grammatical cases in MSA are typically indicated by short vowels, which are commonly omitted from the text. The absence of duals and feminine plurals is also observable, and the inclusion of circumfix negation. In terms of syntax, the prevalence of the verb–subject–object word order is noted to be higher in MSA compared to dialects. Lastly, distinctions in vocabulary are also discernible in the written text.

3 Related Work

In the realm of neural machine translation (NMT) for DA, the predominant emphasis has revolved around translating these dialects into MSA. Nonetheless, a majority of these studies often centre on a singular dialect, as discussed earlier in this document, leading to a deficiency in models that cover a wide range of Arabic dialects.

As an example, Al-Ibrahim and Duwairi (2020), conducted research focusing on translating the Jordanian Arabic dialect into MSA through deep learning techniques, employing an RNN encoder-decoder model. The progress of their work was, however, constrained by the limited size of the corpus.

Likewise, Baniata et al. (2018) addressed the task of translating Levantine dialects, encompassing Jordanian, Syrian, and Palestinian, into MSA. They utilized a comparatively small dataset of parallel sentences sourced from MADAR PADIC corpora. In their approach, they adopted a multitask learning model, where the decoder was shared across various language pairs, while each source language had its dedicated encoder.

In a similar fashion, Kchaou et al. (2022) adopted a hybrid approach in constructing a translation model for the Tunisian dialect. They proposed various augmentation methods to generate a large corpus and subsequently tested different NMT models using this corpus.

In the domain of low-resource NMT for the Algerian Arabic dialect, Hamed et al. (2023), introduced a transductive transfer learning approach. In this approach, the knowledge is conveyed from parent to child models. The evaluation was conducted employing two datasets; MADAR and PADIC. The implementation of the transductive transfer learning approach done by using two types of NMT models namely: Seq2Seq and Attentional-Seq2Seq.

Furthermore, Nagoudi et al. (2022b) developed TURJUMAN [2], a comprehensive neural toolbox with the capability of translating 20 different languages into MSA. The TURJUMAN toolbox leverages the strengths of AraT5 model and explores its proficiency in Arabic decoding. TURJUMAN was developed to utilize semantic similarity for collecting parallel data samples that are openly accessible, ensuring the quality of the collected data. Most recently, researchers in the NMT field, have come to the fact that transfer learning through straightforward fine-tuning is an effective method, particularly when applied between closely related high-resource and low-resource languages (Zoph et al., 2016).

4 Datasets

In Task 2 of OSACT6 Workshop, organizers shared the Dev and Test set in CodaLab [3], for developing and testing purposes respectively. In addition, participants were free to use any of the available linguistics resources and corpora for training their models, called the Training set (Train). Table 1 gives the total number of dialectal sentences in each of the three datasets (train, Dev and Test) used in this research.

Our methodology starts with the training of the chosen AraT5 models, employing five distinct

[2]https://demos.dlnlp.ai/turjuman/

[3]https://codalab.lisn.upsaclay.fr/competitions/17118

datasets, and fine-tuning their hyper-parameters as a result. After the training phase, we evaluated the performance of our models on the Dev set provided by the organizers of the OSACT6 2024 shared task. Ultimately, predictions were generated using the optimal model configuration on the test set. The following sections in this paper will provide details about the selected datasets, AraT5 models, and the training configuration.

Data set	#Sentences
Train	180,211
Dev	1001
Test	1888

Table 1: Number of Sentences in Train, Dev and Test sets

4.1 Train Set:

Shared Task 2 of the OSACT6 Workshop allowed the participants to use any available resources and tools for training and fine-tuning their models. This section details the datasets employed in training our models. While exploring potentially valuable publicly available datasets, we considered those encompassing various Arabic dialects, specifically regional variations pertinent to the five dialects of interest in Shared Task 2 of OSACT6. We identified and made use of five datasets: 1)MADAR, 2)PADIC, 3)Dial2MSA, 4) Arabic semantic textual similarity (STS) and 5)SADID datasets. Table 2 provides statistics regarding the size of each dataset, measured by the number of pairs of DA sentences alongside their corresponding MSA translations.

MADAR (Bouamor et al., 2019) is a parallel corpus that encompasses different Arabic dialects spoken in 25 cities in Arabic world, along with MSA and English. MADAR stands out as the sole corpus in our training data that covers all five dialects of Shared Task 2, namely: Gulf, Egyptian, Levantine, Iraqi, and Maghrebi.

While **PADIC** (Meftouh et al., 2018) is a parallel corpus comprising texts that belong to two primary Arabic dialects alongside the MSA form. It comprises three sub-dialects from Maghrebi: Algerian, Anab, and Tunisian. Additionally, it incorporates two sub-dialects from Levantine: Syrian and Palestinian.

Dial2MSA The Dial2MSA dataset, as outlined by Mubarak (2018), encompasses tweets written in four distinct Arabic dialects: Egyptian, Gulf, Levantine, and Maghrebi, along with their respective MSA translations. It's important to note that the validation process for the translations was carried out manually only for the Egyptian and Maghrebi dialects. In this research, the entire PADIC dataset, which includes translations that have not undergone validation, was employed during the training phase of our models.

Arabic STS dataset collected by, Al Sulaiman et al. (2022), focuses on determining semantic similarity between two given Arabic sentences. Each English phrase was translated into three target languages namely: MSA, Egyptian and Saudi dialect

SADID (Abid, 2020) is a parallel corpus for English, Egyptian, Levantine and MSA. The dialectal texts were collected from three distinct sources:1) Wikipedia for its diverse domains and clear language, 2) Aesop's Fables for its narrative style, and 3) specific dialogues from movie subtitles. English was chosen as the source language for sentences rather than MSA to avoid introducing bias into the translations (Bouamor et al., 2014). Various translators offer translations with varying degrees of dialectal influence.

	Glf	**Egy**	**Lev**	**Iraqi**	**Magh**
MADAR	15400	13800	18600	18600	29200
PADIC	0	0	12824	0	19236
Dial2MSA	18010	16355	18000	0	7912
Arabic STS	2758	2758	0	0	0
SADID	0	2997	2997	0	0
Total	36168	35910	52421	18600	37112

Table 2: The number of dialect-to-MSA translation sentences in each of the datasets used in Task 2

4.2 Dev Set

The development set[4] is structured as a JSON file, containing 1001 sentences, with approximately 200 sentences allocated to each dialect. This dataset is essential for improving and evaluating translation systems, with a focus on achieving outstanding results. As you can see in the figure 2, each sentence in the development set has a unique identifier ("id"). The second key is the dialect name label ("dialect"), to which the sentence belongs. The third key in the dictionary is ("source"), representing the textual content of the sentence. Additionally, the key ("target") contains the translation of the sentence into MSA.

[4] https://osact-lrec.github.io

Figure 1: Capture of the JSON File Structure for the Dev Set

4.3 Test Set

The test set[5] is structured as a JSON file, containing a total of 1888 sentences, with approximately 377 sentences allocated to each dialect. These test sentences have been carefully crafted to evaluate the performance of translation systems to accurately convert DA text into MSA. As you can see in Figure 2, each sentence in the test set has a unique identifier ("id"). The second key is the dialect name label ("dialect"), to which the sentence belongs. The third key in the dictionary is ("source"), representing the textual content of the sentence.

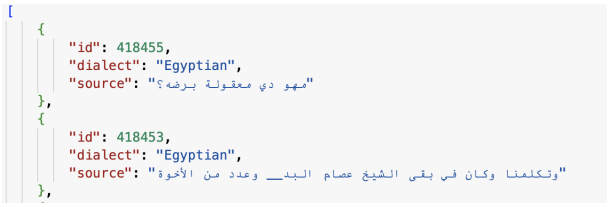

Figure 2: Capture of the JSON File Structure for the Test Set

5 Methodology

Within this section, we present the AraT5 models that serve as our foundation, illustrate the fine-tuning process, and delve into the optimization of hyper-parameters.

5.1 Training Configurations

From the train set, we have observed that the dialectal text and the corresponding MSA text share the same words between them. Based on this observation, we have applied the same method as (Khered et al., 2023), this involves generating an additional pair for every translation pair in our Train set, in which both the source and the target consist of text written in MSA. Table 3 shows an example of the additional pair generation in the Train set. Leveraging these additional pairs empowers our models to grasp the nuances of sentences containing words shared with MSA. In our training setup, we've incorporated all dialect-to-MSA translation pairs from the Train set, focusing on regions pertinent to the five targeted dialects used in training a single model. Consequently, translation pairs from datasets covering Gulf, Egyptian, Levantine, Iraqi, and Maghrebi dialects were employed in the model learning process.

Source	Target
Original Pair	
رجال ياكل مكرونة	رجل يأكل المعكرونة
Additional Pair	
رجل يأكل المعكرونة	رجل يأكل المعكرونة
English Translation	
A man is eating pasta	

Table 3: An example of adding MSA pair to the Train set in which, the source and target are both the MSA translation of the source text

5.2 Fine-Tuning AraT5 Models

The Text-To-Text Transfer Transformer (T5) model transforms various natural language processing (NLP) tasks into a consistent textual format. Among the NLP tasks on which T5 has been pre-trained is MT (Raffel et al., 2020). In our study, we conducted fine-tuning on four distinct AraT5 models: AraT5 base, AraT5v2-base-1024, AraT5-MSA-Small, and AraT5-MSA-Bases

- The **AraT5 base** This model by Nagoudi et al. (2022a), is a tailored version of T5, meticulously fine-tuned to handle and process Arabic text. Functioning as a fundamental model, It demonstrates versatility across a range of natural language processing tasks, including text classification, text generation, and machine translation (MT). AraT5-base effectively leverages the Transformer architecture and pre-trained embeddings to understand and generate Arabic text proficiently.
- The **AraT5v2-base-1024** model signifies an advanced version of AraT5-Base. In the latest iteration of AraT5, AraT5v2[6], the sequence length has been expanded from 512 to 1024, this represented as "1024" in its name. This

[5] https://osact-lrec.github.io

[6] https://huggingface.co/UBC-NLP/AraT5v2-base-1024

extended sequence length significantly enhances the model's adaptability across various NLP tasks. Notably, the fine-tuning process of AraT5v2-base-1024 demonstrates convergence approximately 10 times faster than its predecessor, AraT5-base. This accelerated convergence has the potential to considerably expedite both training and fine-tuning procedures, thereby improving overall efficiency. The selection of this model to be included in our experiments stems from its outstanding performance, as illustrated in Table 5, where its performance surpassed that of other models.

- The **AraT5-MSA-Base** (Nagoudi et al., 2022a), represents an enhanced iteration of AraT5, specifically designed to proficiently handle diverse standard Arabic natural language processing tasks. With an augmented architecture and an increased number of parameters, it excels in tackling intricate tasks that require a profound understanding of the language. AraT5-MSA-Base stands out as an ideal choice for research projects and applications demanding advanced linguistic modelling.

- In contrast **AraT5- MSA-Small** (Nagoudi et al., 2022a), is a refined iteration of the AraT5 model, known as AraT5-MSA-Small, is specifically designed for the streamlined processing of Modern Standard Arabic (MSA) data. It operates at an accelerated pace and requires fewer computational resources compared to its "Base" counterpart. AraT5-MSA-Small is commonly utilized in applications where operational efficiency is crucial, all without a substantial sacrifice in quality.

Our methodology encompassed fine-tuning the above mentioned models using the whole Train set together with all of the four selected AraT5 models. Moreover, the same hyper-meters being used for fin-tuning the models. This standardized methodology empowered us to conduct significant comparisons between the models' performance in our experiments. Table 4 provides information about the hyper-meters used during the training process. All the models employed in our experiments were obtained from the Hugging Face[7] repository. The PyTorch Transformers library[8] is used for designing and executing our Python codes. These hyper-parameters were meticulously chosen to attain optimal performance while reducing the duration of training

Parameters	Values
learning_rate	5e-5
max_target_length	128
max_source_length	128
per_device_train_batch_size	16
per_device_eval_batch_size	16
save_steps	1000
eval_steps	1000
num_train_epochs	2

Table 4: Hyper-parameters for fin-tuning the AraT5 models

6 Results and Discussion

All models utilized in our research underwent evaluation using the BiLingual Evaluation Understudy (BLEU) metric (Papineni et al., 2002), which measures the matching between text generated by the machine (model) and the reference translation based on overlapping words.Table 5 presents the evaluation of model performance, measured in terms of BLEUscore. Notably, the AraT5v2-base-1024 model stands out as the top-performing model, achieving overall BLEUscore of 21.0 when used on the Dev set.

The performance evaluation on the chosen AraT5 models and learning hyper-parameters underscores the intricacies of the translation task, particularly in translating from AD to MSA.The low BLEUscores in our experiments can be attributed to various factors. These encompass issues in the availability of corpora for some dialects in this study, notably the small size of the Iraqi dialect in the total Train set. Additionally, due to time and computational resource constraints,we could not investigate the the impact of the values of varying hyperparameter on the AraT5 models' performance. These combined factors pose challenges in obtaining higher performance reults in Arabic MT tasks. Enhancement of existing resources and creation of new comprehensive Arabic parallel datasets will lead to improvement in the translation outcomes in the future.

[7]https://huggingface.co

[8]https://pytorch.org/hub/huggingface_pytorch-transformers/

Model	BLEUScore
AraT5 base	19.26
AraT5v2-base-1024	21.0
AraT5-MSA-Base	16.88
AraT5- MSA-Small	15.97

Table 5: BLEUscores on the Dev set of the chosen models.

7 Conclusion

This paper outlines our contributions to the OSACT6 2024 Shared Task 2, which revolves around MT of AD into MSA using five Arabic parallel datasets: MADAR, PADIC, Dial2MSA, Arabic STS, and SADID. Throughout our research, we examined four variants of the AraT5 model: AraT5 base, AraT5v2-base-1024, AraT5-MSA-Small, and AraT5-MSA-Base. The experimental findings presented in this study suggest the potential application of these methods to automate the construction of Arabic parallel corpora. Moreover, our commitment extends to advancing research through additional exploration of fine-tuning techniques for transformer models.

Potential future directions include the development of a multilingual model tailored to DA and MSA. Another avenue involves the creation of additional Arabic parallel corpora covering underresourced Arabic dialects, for example, a corpus of Saudi regional dialects. Additionally, the prevalence of Arabizi—where young Arabs on social media use the Latin script and numerals to represent Arabic sounds—represents an important phenomenon to consider for future research endeavours.

References

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wael Abid. 2020. The sadid evaluation datasets for low-resource spoken language machine translation of arabic dialects. In *International Conference on Computational Linguistics*.

Roqayah M. Al-Ibrahim and Rehab Duwairi. 2020. Neural machine translation from jordanian dialect to modern standard arabic. *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 173–178.

Mahmoud A Al-Sobh, Abdel-Rahman H Abu-Melhim, and Nedal A Bani-Hani. 2015. Diglossia as a result of language variation in arabic: Possible solutions in light of language planning. *Journal of Language Teaching and Research*, 6(2):274.

Mansour Al Sulaiman, Abdullah M. Moussa, Sherif Abdou, Hebah Elgibreen, Mohammed Faisal, and Mohsen Rashwan. 2022. Semantic textual similarity for modern standard and dialectal arabic using transfer learning. *PLOS ONE*, 17(8):1–14.

Laith H. Baniata, Seyoung Park, and Seong-Bae Park. 2018. A multitask-based neural machine translation model with part-of-speech tags integration for arabic dialects. *Applied Sciences*, 8(12).

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2023. Investigating lexical replacements for Arabic-English code-switched data augmentation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 86–100, Dubrovnik, Croatia. Association for Computational Linguistics.

Saméh Kchaou, Rahma Boujelbane, Emna Fsih, and Lamia Hadrich-Belguith. 2022. Standardisation of dialect comments in social networks in view of sentiment analysis : Case of Tunisian dialect. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5436–5443, Marseille, France. European Language Resources Association.

Abdullah Khered, Ingy Abdelhalim, Nadine Abdelhalim, Ahmed Soliman, and Riza Batista-Navarro. 2023. UniManc at NADI 2023 shared task: A comparison of various t5-based models for translating Arabic dialectical text to Modern Standard Arabic. In *Proceedings of ArabicNLP 2023*, pages 658–664, Singapore (Hybrid). Association for Computational Linguistics.

K. Meftouh, S Harrat, and Kamel Smaïli. 2018. PADIC: extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey.

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. *OSACT*, 3:49.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022a. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022b. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 1–11, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Houcemeddine Turki, Emad Adel, Tariq Daouda, and Nassim Regragui. 2016. A conventional orthography for maghrebi arabic. In *International Conference on Language Resources and Evaluation*.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.